DOI:
10.3389/fpsyg.2015.01750

Document Version
Publisher's PDF, also known as Version of record

frontiers in Psychology

MINI REVIEW
published: 12 November 2015
doi: 10.3389/fpsyg.2015.01750

What is the Value of Embedding Artificial Emotional Prosody in Human–Computer Interactions? Implications for Theory and Design in Psychological Science

Rachel L. C. Mitchell[1]* *and Yi Xu*[2]

[1] Centre for Affective Disorders, Institute of Psychiatry Psychology and Neuroscience, King's College London, London, UK, [2] Speech Hearing and Phonetic Sciences, Division of Psychology and Language Sciences, University College London, London, UK

In computerized technology, artificial speech is becoming increasingly important, and is already used in ATMs, online gaming and healthcare contexts. However, today's artificial speech typically sounds monotonous, a main reason for this being the lack of meaningful prosody. One particularly important function of prosody is to convey different emotions. This is because successful encoding and decoding of emotions is vital for effective social cognition, which is increasingly recognized in human–computer interaction contexts. Current attempts to artificially synthesize emotional prosody are much improved relative to early attempts, but there remains much work to be done due to methodological problems, lack of agreed acoustic correlates, and lack of theoretical grounding. If the addition of synthetic emotional prosody is not of sufficient quality, it may risk alienating users instead of enhancing their experience. So the value of embedding emotion cues in artificial speech may ultimately depend on the quality of the synthetic emotional prosody. However, early evidence on reactions to synthesized non-verbal cues in the facial modality bodes well. Attempts to implement the *recognition* of emotional prosody into artificial applications and interfaces have perhaps been met with greater success, but the ultimate test of synthetic emotional prosody will be to critically compare how people react to synthetic emotional prosody vs. natural emotional prosody, at the behavioral, socio-cognitive and neural levels.

Keywords: social cognition, emotion, prosody, artificial speech, human–computer interaction

OPEN ACCESS

Edited by:
Alessandro Vinciarelli, University of Glasgow, UK

Reviewed by:
Gelareh Mohammadi, University of Geneva, Switzerland
Ronald Böck, Otto von Guericke University Magdeburg, Germany

***Correspondence:**
Rachel L. C. Mitchell
rachel.mitchell@kcl.ac.uk

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *24 July 2015*
Accepted: *31 October 2015*
Published: *12 November 2015*

Citation:
Mitchell RLC and Xu Y (2015) What is the Value of Embedding Artificial Emotional Prosody in Human–Computer Interactions? Implications for Theory and Design in Psychological Science. Front. Psychol. 6:1750. doi: 10.3389/fpsyg.2015.01750

INTRODUCTION

One of the great challenges faced by the human mind is the need to comprehend the mental state of other people. Fortunately, this task is made easier by non-verbal cues in the form of emotional prosody, and it is these such cues that people use to manage their social relationships (Mehu and Scherer, 2012; Tschacher et al., 2014). Prosody refers to acoustic properties beyond those of consonants and vowels, including variables such as pitch, duration, intensity, voice quality, and spectral properties (Ross, 2010). By manipulating prosody, we can alter our tone of voice, and hence change the emotion conveyed. Human–computer interaction concerns how people interact with computerized technology (Boehm-Davis, 2008), and amongst the range of applications and

interfaces (HCI-AI) that exist, an ever increasing proportion incorporate artificial speech (Drahota et al., 2008; Robinson and el Kaliouby, 2009). However, a likely obstacle to wide acceptance of today's artificial speech is its lack of the "human touch," because it sounds monotonous, and does not change interactively with the user. Whilst an increasing number of HCI-AI now also incorporate rudimentary speech recognition (Putze and Schultz, 2014; Schwenker et al., 2015), they too often lack the "human touch" because they cannot distinguish between the words that are spoken and the way in which they are spoken. In the following mini-review we consider the possible advantages and disadvantages of incorporating expressive speech with emotional prosody into HCI-AI. We examine what has been achieved so far, and the necessary work that remains.

WHEN IS ARTIFICIAL EMOTIONAL PROSODY BENEFICIAL FOR USERS?

It has been known for some time that the ability to recognize and express emotions plays a key role in human communication, and more recently, its importance has been recognized in HCI. In research on HCI this is reflected in the wish to achieve more natural interaction (Schröder, 2001), dubbed the "Realism Maximization Theory" (Edlund et al., 2008) Naturally, realism maximization cannot be achieved without incorporating emotional prosody into HCI technology to enhance the communication of intended messages, just as these cues do for natural speech (Esposito et al., 2015). Whilst human and artificial voices can sometimes be distinguished (Belizaire et al., 2007; Gaucher et al., 2013), available evidence suggests that whether synthesized or recorded, a happy voice still makes content seem happier, and a sad voice still makes content seem less happy (Nass et al., 2001).

The key question is, has the aim of enhancing HCI by incorporating synthesized emotional prosody been achieved? Engagement with HCI continues to increase because it is believed that the experience is acceptable and enjoyable, and that HCI-AI can serve as socially intelligent interaction partners that can provide assistance to people (Gorostiza and Salichs, 2011; Wood et al., 2013). But is this really the case? One forum in which HCI-AI appears to be benefitting from incorporating emotional prosody is in the healthcare system. Here, communications to and from patients through emotional channels is of vital importance, and advocates argue that these technologies make HCI-AI more human-like, meaning that users can then rely on well-learned social interaction skills to make the interactions smoother (Berry et al., 2005). Augmentative and alternative communication devices for those unable to produce their own speech (e.g., because of neurodegenerative disease) are a particularly relevant example of the benefits of adding emotional prosody to HCI-AI. As illustrated in the case of the eminent Stephen Hawking, with these devices users can spell out or select any word they choose, but only have recourse to punctuation to influence the way those words are spoken. When listening to users of these devices other than Hawking, listeners may incorrectly assume that they are emotionally as well as speech-impaired, and socially inept (Pullin and Cook, 2013). This is unsatisfactory given the evidence that without the full embodiment of emotional expression present in interpersonal interactions, communication, coordination and performance can suffer immensely (Gerdes et al., 2014). Other applications of expressive speech are covered elsewhere in this review, but for further examples of areas of application which appear promising at this stage, we direct the reader to consider narratives in gaming arenas, voice conversion for the purposes of, and the use of human–human interaction analyses to aid our understanding of social dynamics (Burkhardt and Stegmann, 2009; Schröder, 2009; Vinciarelli et al., 2009; Devillers and Campbell, 2011; Creer et al., 2013; Esposito et al., 2015).

Compared to the expression of emotional prosody, there appears to be greater evidence of success in the incorporation of prosodic emotion recognition into HCI-AI. The driving force here has been the need for these applications to make appropriate reactions in interactive processes, for which the capacity to process human speech signals through emotion recognition is required (Wang et al., 2014; Mavridis, 2015). Whilst many potential applications are being developed (Burkhardt and Campbell, 2014). In the healthcare context, people with autism can sometimes prefer to communicate with computers rather than humans, because it feels more predictable to them and affords greater control over an otherwise chaotic social world (Moore et al., 2000). Sociable HCI-AI incorporating microphones to record emotional prosody may be a good approach to helping social interaction skills (Kim et al., 2013). In this arena, affect sensing and recognition technologies can help increase self-awareness, and provide novel means of self-monitoring (el Kaliouby et al., 2006). Another possible healthcare application is to improve diagnoses of flat affect in patients with depression and schizophrenia, which currently rely on psychiatrists' subjective judgment (Fragopanagos and Taylor, 2005). Similarly, there is evidence of the promise of automated analyses of vocal markers for Parkinson's disease (Tsanas et al., 2012). HCI-AI sensitive to emotional prosody also holds promise in the learning environment, e.g., in automatic tutoring applications. Such emotion-sensitive automatic tutors have the capacity to interactively adjust teaching content and the speed at which it is delivered, based on whether a user finds it boring and dreary, or exciting and thrilling (Litman and Forbes-Riley, 2006). Whilst such systems were once feared not to be as effective as one-to-one human tutoring, the addition of the capacity to recognize emotional signals, has narrowed the gap (Mao and Li, 2010).

WHEN IS IT NOT BENEFICIAL?

Unfortunately, the use of artificial speech technology that does not deliver its promise in terms of improved interaction will only frustrate users (Laukka et al., 2011). At the theoretical level, Mori's "uncanny valley" theory (Edlund et al., 2008) suggests that as artificial HCI characters approach realistic visual similarity to humans, at a certain point they stop being likeable and instead appear eerie, frightening, repulsive—"uncanny" (Mori, 1970). What if the same "uncanny" effect were true for auditory dimensions of HCI-AI, and how might such an auditory effect compromise user acceptance? Certainly disembodied emotional voices presented in isolation are not always received well. Here

users may struggle to interpret emotional cues conveyed through the voice, because unlike embodied HCI-AI voices that possess contextual cues to allow users to better determine emotional intentions, disembodied voices allow too much ambiguity (Barker et al., 2013). On a practical level, some question how feasible the development of human-like artificial prosody truly is (Edlund et al., 2008). In particular, there is no singular means of creating artificial prosody, and each specific means has its own imperfections (Esposito and Esposito, 2012). Ultimately, however, the issue boils down to the question of whether and to what extent artificial emotion cues are contextually appropriate, which in turn will be contingent on our level of scientific understanding of emotional expressions.

Questions have also arisen as to the success with which artificial emotional prosody has usefully been incorporated into HCI-AI, in some of the very same areas that heralded its promise. To take the example of ATM's with emotional prosody, compared to human tellers or traditional ATMs without this capacity, it could be argued that it is a rather unusual experience to talk to a machine (Fischer, 2010). This might be due, in part, to the awkwardness of knowing that no human is there, and it might also be difficult to imagine that for ordinary use, people want to have an artificial agent openly express anger or displeasure to a user. Or it might simply reflect that this technology is relatively new, and that consumers and users need time to get used to such well-intentioned amplification of emotional communication through prosody. A second exemplar concern is the healthcare applications for those unable to produce their own speech. Here there is no vocal individuality, i.e., the systems are not designed to imitate a specific speaker's voice (Wendemuth and Biundo, 2011). This identity mismatch may impact use and adoption of these devices and further perpetuate the divide between the user and the device (Mills et al., 2014).

WHAT ABOUT THE FUTURE?

If it were possible to further improve the quality of artificial speech with emotional prosody, it would have significant consequences for those involved in creating HCI-AI (Brenton et al., 2005). As computerized technology becomes an ever greater fixture at home and at work, our future interactions with it will need to become even more sophisticated (Wendemuth and Biundo, 2011; Honold et al., 2014). Some time ago, it was recommended that artificial speech synthesis technology should not only have the ability to control prosody based on meaning, but also the capability to control individual speaking style (another form of prosody), choosing application-oriented speaking styles, and be able to add emotion (Furui, 1995). Yet, as we have seen, there remains much work to be done (Burkhardt and Stegmann, 2009). Although problematic, delivery to date of HCI-AI able to crudely interact with people, attempt to sense emotional prosody, and try to produce suitable responses, has produced great expectations for the future (Esposito and Esposito, 2012).

Social HCI-AI need to follow behaviors similar to humans: they interact and communicate with humans by following a set of social rules (Pullin and Cook, 2013). Additional work on the social skills and responsivity with which HCI-AI are programmed will likely increase the empathy and acceptance level of interactions further (Leite et al., 2013). From the human interface point of view, it has long been recognized that HCI-AI should be able to automatically acquire new knowledge about the thinking process of individual users, automatically correct user errors, and understand user intentions by accepting rough instructions and inferring details (Furui, 1995). Ultimately, the hope for the future is that HCI-AI could extract the prosodic cues from a user's speech, capitalize on the information to inform predictive models of likely emotions (Litman and Forbes-Riley, 2006), and amend their own displays and actions accordingly. Such an aim is not without its challenges though. For example, much work is ongoing at present into how a HCI-AI might best transcribe and annotate a user's prosodic emotion cues in order to reliably label and act appropriately on the likely emotional state conveyed thus (Siegert et al., 2013, 2014). Beyond being responsive and interactive, HCI-AI with emotional prosody also requires further work on the modification of their implementation depending on context. We may adopt a different palette of tones of voice with different people, depending on our relationship to them or the social context (Pullin and Cook, 2013).

Whilst there may be problems with current technology (Schröder, 2009), we believe that healthcare applications of HCI-AI with artificial emotional prosody still hold the potential to make a genuine difference to peoples' lives in the future. HCI-AI that express emotion cues could especially enhance the ability to make sense of and communicate with others in people with difficulty understanding, communicating and regulating their emotion systems, such as autism, and the affective disorders (Robinson and el Kaliouby, 2009). Personalized voices may also be possible for those who rely on alternative or augmented communication devices, by mapping existing text to speech corpora onto a voice personalized to the residual vocal characteristics of a specific user (Creer et al., 2013; Mills et al., 2014). Whilst future assistive HCI-AI agents with emotional prosody might be expected to benefit users of all ages, the development of socially intelligent assistive technology has promise for increasing quality of life for older adults in particular, in the form of reminder systems, telecommunication systems, surveillance systems, and the ability to provide social interaction and complete daily household tasks (Beer et al., 2015).

PROBLEMS STILL REQUIRING A SOLUTION

As alluded to above, current technology for synthesizing effective emotional prosody is still rudimentary. Yet if we were able to improve its quality, it would have far-reaching consequences for the success of HCI-AI attempting to capitalize on its potential (Brenton et al., 2005). One source of the slow progress toward synthesizing good quality artificial emotional prosody is the difficulty of identifying clear acoustic correlates for discrete emotions (Banse and Scherer, 1996; Schröder, 2009). This problem is compounded by the fact that most work on natural emotional prosody has taken its measurements from recordings of actors trying to portray various emotional tones of voice. But it is questionable whether actors' portrayals authentically represent the characteristics of speech used by ordinary people

when they spontaneously experience emotions (Douglas-Cowie et al., 2007; Esposito and Esposito, 2012). A further difficulty is that, because each acoustic dimension can be measured in many different ways (Xu, 2011), there are actually many more possible acoustic measurements than can be realistically exhausted in elucidating acoustic correlates. However, recent work on defining emotions through mathematical modeling holds promise in tackling the current lack of consistent acoustic correlates in human communications (Hartmann et al., 2012; Xu, 2015).

A second problem with generating good quality artificial emotional prosody is that there is still an unsolved need to directly control specific aspects of artificial prosodic emotion cues based on theoretical motivations, as attempted in dimensional approaches to the measurement of emotions (Burkhardt and Stegmann, 2009; Mauss and Robinson, 2009; Verma and Tiwary, 2014). However, effective control methods for empirically investigating vocal emotional expressions have not yet been developed, and although some links have been found between activation (arousal) and parameters such as pitch and intensity, F_0 range, articulation rate etc., no acoustic parameters have been identified as reliable indicators of key dimensions of emotion such as valence or approach/avoidance (Mauss and Robinson, 2009). To address the lack of theoretical foundation, robust algorithmic implementation of situated social information processing facets is necessary, and it is likely that multiple theoretical perspectives will need to be considered, ranging from mathematical models and dynamics of signal exchanges (e.g., emotional states and context effects), to social intelligence, behavioral analyses, and cognitive processes such as cooperation (Esposito et al., 2015).

An interesting recent development is an ethological approach to emotional prosody, which examines commonalities been animal calls and human emotional prosody (Xu et al., 2013a). The theoretical framework was first developed in a study of animal calls (Morton, 1977) and later extended to humans (Ohala, 1984). It posits a strong selection pressure for organisms to vocally (just as they do visually) manipulate their apparent body size when interacting with others. This size-projection hypothesis has been shown to be capable of explaining a broad range of animal and human behaviors, as well as bodily anatomies, including the acoustic characteristics of animal calls (Fitch and Kelley, 2000; Reby and McComb, 2003; Reby et al., 2005; Harris et al., 2006; Charlton et al., 2007, 2011), the descent or elongation of the vocal tract (Fitch, 1999; Fitch and Reby, 2001), and sexually and socially related human vocal behavior (Feinberg et al., 2005, 2006, 2008; Bruckert et al., 2006; Riding et al., 2006; Fraccaro et al., 2011; Xu et al., 2013b). The relevance of this hypothesis for human emotional prosody has been demonstrated by a series of studies (Chuenwattanapranithi et al., 2008; Noble and Xu, 2011; Xu et al., 2013a,b). The consistency of results shown in these studies suggest that the ethological-based approach is promising and needs to be further explored in future research.

EVALUATION OF SUCCESS

In this review, we have seen that the ability to synthesize emotional prosody might, for the most part, be desirable (Drahota et al., 2008; Robinson and el Kaliouby, 2009). As work continues, it will be important to understand and properly evaluate our predispositions to artificial emotional prosody (Robinson and el Kaliouby, 2009). It is particularly vital to know whether these cues are perceived in the same way as human emotion cues. After all, the most critical test is to put artificial speech technology into action and to expose it to the critical comparison with social reality as created by nature (Vogeley and Bente, 2010). To make such a judgement, we believe that a wide array of assessments will be required.

At the behavioral level, some literature suggests people may not be as good at identifying synthesized facial expressions in avatars as they are at identifying human expressions (Moser et al., 2007; Rosset et al., 2008), perhaps because many are clearly not perceptually natural (Douglas-Cowie et al., 2007; Cowie, 2009). However, other studies have shown that recognition of synthesized facial expressions can sometimes match or even surpass that of human expressions (Dyck et al., 2008). Whilst human and artificial voices can be distinguished at the behavioral level (Nass et al., 2001; Belizaire et al., 2007), a clear main effect has not been demonstrated when judging intonational emotions. Another complication comes from the fact that facial studies report that identification of emotion from artificial faces may be emotion-dependent, and that recognition is worst for disgust (Moser et al., 2007; Dyck et al., 2008). The uncanny valley theory might more generally predict that negative valence emotions (aversive or warning stimuli) such as anger, fear, sadness and disgust might attract lower ratings of familiarity and human-likeness (Tinwell et al., 2011). The validity of such predictions has never been tested for artificial emotional prosody.

Of particular relevance at the socio-cognitive level, would be to determine what criteria afford a machine the status of "social agent" (a software agent or robot capable of social communication with human beings; Aharoni and Fridlund, 2007). Despite the knowledge that computerized technology does not warrant social treatment, people nonetheless tend to apply social expectations to and exhibit the same responses to computerized technology as they would to human communication partners (Lee, 2010). Indeed, there is prior evidence for impression formation from artificial emotional prosody. For example, its implementation has been shown to influence social judgments of liking and credibility from synthesized speech (Nass et al., 2001), and people tend to present themselves in a more positive light to HCI agents that emit artificial speech (Parise et al., 1999), although findings of impression management in response to artificial speech are not as strong as that in response to human speech (Lee, 2010; Mitchell et al., 2011). We suggest that the social influence of HCI-AI agents with artificial emotional prosody may even extend to stereotypical impression formation. Indeed, the literature on "Sensitive Artificial Listeners" illustrates that it is possible for two people to have a conversation in which one pays little or no attention to the meaning of what the other says, and chooses emotional interpretations and subsequent responses on the basis of quite superficial cues(Douglas-Cowie et al., 2008). Recent data has even shown that participants might prefer robots with matching gender-occupational role stereotypes (Tay et al., 2014), and it would be easy enough to test the hypothesis that emotional prosody makes synthesized speech

more human-like, and therefore more susceptible to human-like stereotypes.

In assessing the utility of artificial emotional prosody at the neural level, the activation of "social cognition" brain regions is more significant for human stimuli than for artificial stimuli, as if our brains are "tuned in" to the former (Vogeley and Bente, 2010). Most pertinent, this is observed when participants derive emotion cues from facial expressions, in the amygdala, insula and prefrontal cortex (Mullennix et al., 2003; Moser et al., 2007; Cheetham et al., 2011; de Borst and de Gelder, 2015). Thus it could be possible to use the activation intensity of such brain regions to assess how human-like an artificial emotion stimulus is (Chaminade and Cheng, 2009). With respect to prosody, researchers should also measure the activation of regions associated with "voiceness" (degree to which an auditory stimulus resembles the human voice) i.e., bilateral upper banks of the superior temporal sulcus (Belizaire et al., 2007), and those associated with prosodic emotion recognition, i.e., the right posterior middle/superior temporal gyri (Mitchell et al., 2003). We would recommend that the neural response of emotion-specific regions such as the amygdala for fear (Janak and Tye, 2015), insula for disgust (Chapman and Anderson, 2012), and superior temporal sulcus for anger (Carter and Pelphrey, 2008) should also be probed in evaluating artificial emotional prosody. However, it needs to be borne in mind that there is no singular means of creating artificial prosody. Thus each specific means of synthesizing emotional prosody will have its own imperfections, its own acoustic correlates, and might invoke its own pattern of neural response.

CONCLUDING THOUGHTS

Given the importance of emotional prosody in human to human communication, there is significant potential for interactions between humans and computerized technology to benefit from including synthesized emotional prosody in HCI-AI. This need is only amplified by the pace with which HCI-AI are evolving. Indeed, to quote Picard "If we want computers to be genuinely intelligent and to interact naturally with us, we must give computers the ability to recognize, understand, and even to have and express emotions" (Picard, 1997). Whilst HCI has mostly been enhanced by including artificially synthesized *facial* expressions, a full multi-level evaluation of our reactions to synthesized emotional prosody is needed before the wisdom of its inclusion can be properly evaluated. Further work is also necessary to determine whether there are circumstances in which its inclusion does not work well, or whether the problem lies in how it is synthesized. Achieving human-likeness dialogs with HCI-AI through explicit computational models might also provide valuable insights about how humans communicate with each other (Edlund et al., 2008).

REFERENCES

Aharoni, E., and Fridlund, A. J. (2007). Social reactions toward people vs. computers: how mere labels shape interactions. *Comput. Hum. Behav.* 23, 2175–2189. doi: 10.1016/j.chb.2006.02.019

Banse, R., and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol.* 70, 614–636. doi: 10.1037/0022-3514.70.3.614

Barker, P., Newell, C., and Newell, G. (2013). Can a computer-generated voice be sincere? A case study combining music and synthetic speech. *Logoped. Phoniatr. Vocol.* 38, 126–134. doi: 10.3109/14015439.2013.795605

Beer, J. M., Smarr, C. A., Fisk, A. D., and Rogers, W. A. (2015). Younger and older users' recognition of virtual agent facial expressions. *Int. J. Hum. Comput. Stud.* 75, 1–20. doi: 10.1016/j.ijhcs.2014.11.005

Belizaire, G., Fillion-Bilodeau, S., Chartrand, J. P., Bertrand-Gauvin, C., and Belin, P. (2007). Cerebral response to 'voiceness': a functional magnetic resonance imaging study. *Neuroreport* 18, 29–33. doi: 10.1097/WNR.0b013e3280122718

Berry, D. C., Butler, L. T., and De Rosis, F. (2005). Evaluating a realistic agent in an advice-giving task. *Int. J. Hum. Comput. Stud.* 63, 304–327. doi: 10.1016/j.ijhcs.2005.03.006

Boehm-Davis, D. A. (2008). Discoveries and developments in human–computer interaction. *Hum. Factors* 50, 560–564. doi: 10.1518/001872008X288529

Brenton, H., Gillies, M., Ballin, D., and Chatting, D. J. (2005). The uncanny valley: does it exist? *Paper Presented at the 11th International Conference on Human–Computer Interaction*, Las Vegas, NV, USA.

Bruckert, L., Lienard, J. S., Lacroix, A., Kreutzer, M., and Leboucher, G. (2006). Women use voice parameters to assess men's characteristics. *Proc. Biol. Sci.* 273, 83–89. doi: 10.1098/rspb.2005.3265

Burkhardt, F., and Campbell, N. (2014). "Emotional speech synthesis," in *The Oxford Handbook of Affective Computing*, eds R. Calvo, S. D'Mello, J. Gratch, and A. Kappas (New York: Oxford University Press), 286.

Burkhardt, F., and Stegmann, J. (2009). "Emotional speech synthesis: applications, history and possible future," in *Proceedings of Electronic Speech Signal Processing,* Deutsche Telekom Laboratories, Berlin.

Carter, E. J., and Pelphrey, K. A. (2008). Friend or foe? Brain systems involved in the perception of dynamic signals of menacing and friendly social approaches. *Soc. Neurosci.* 3, 151–163. doi: 10.1080/17470910801903431

Chaminade, T., and Cheng, G. (2009). Social cognitive neuroscience and humanoid robotics. *J. Physiol. Paris* 103, 286–295. doi: 10.1016/j.jphysparis.2009.08.011

Chapman, H. A., and Anderson, A. K. (2012). Understanding disgust. *Ann. N. Y. Acad. Sci.* 1251, 62–76. doi: 10.1111/j.1749-6632.2011.06369.x

Charlton, B. D., Ellis, W. A. H., Mckinnon, A. J., Cowin, G. J., Brumm, J., Nilsson, K., et al. (2011). Cues to body size in the formant spacing of male koala (*Phascolarctos cinerus*) bellows: honesty in an exaggerated trait. *J. Exp. Biol.* 214, 3414–3422. doi: 10.1242/jeb.061358

Charlton, B., Reby, D., and Mccomb, K. (2007). Female perception of size-related formant shifts in a nonhuman mammal. *Anim. Behav.* 74, 707–714. doi: 10.1016/j.anbehav.2006.09.021

Cheetham, M., Suter, P., and Jancke, L. (2011). The human likeness dimension of the "uncanny valley hypothesis": behavioral and functional MRI findings. *Front. Hum. Neurosci.* 5:126. doi: 10.3389/fnhum.2011.00126

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., and Maneewongvatana, S. (2008). Encoding emotions in speech with the size code: a perceptual investigation. *Phonetica* 65, 210–230. doi: 10.1159/000192793

Cowie, R. (2009). Perceiving emotion: towards a realistic understanding of the task. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 3515–3525. doi: 10.1098/rstb.2009.0139

Creer, S., Cunningham, S., Green, P., and Yamagishi, J. (2013). Building personalised synthetic voices for individuals with severe speech impairment. *Comput. Speech Lang.* 27, 1178–1193. doi: 10.1016/j.csl.2012.10.001

de Borst, A. W., and de Gelder, B. (2015). Is it the real deal? Perception of virtual characters versus humans: an affective cognitive neuroscience perspective. *Front. Psychol.* 6:576. doi: 10.3389/fpsyg.2015.00576

Devillers, L., and Campbell, N. (2011). Special issue of computer speech and language on "affective speech in real-life interactions." *Comput. Speech Lang.* 25, 1–3. doi: 10.1016/j.csl.2010.07.002

Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., and Heylen, D. (2008). "The Sensitive Artificial Listner: an induction technique for generating emotionally coloured conversation," in *LREC Workshop on Corpora for Research on Emotion and Affect,* (Marrakech, Marokko), 1–4.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., et al. (2007). "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," in *Affective Computing and Intelligent*

Interaction, eds A. R. Paiva, R. Prada, and R. Picard (Heidelberg: Springer Berlin), 488–500.

Drahota, A., Costall, A., and Reddy, V. (2008). The vocal communication of different kinds of smile. *Speech Commun.* 50, 278–287. doi: 10.1016/j.specom.2007.10.001

Dyck, M., Winbeck, M., Leiberg, S., Chen, Y., Gur, R. C., and Mathiak, K. (2008). Recognition profile of emotions in natural and virtual faces. *PLoS ONE* 3:e3628. doi: 10.1371/journal.pone.0003628

Edlund, J., Gustafson, J., Heldner, M., and Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Commun.* 50, 630–645. doi: 10.1016/j.specom.2008.04.002

el Kaliouby, R., Picard, R., and Baron-Cohen, S. (2006). Affective computing and autism. *Ann. N. Y. Acad. Sci.* 1093, 228–248. doi: 10.1196/annals.1382.016

Esposito, A., and Esposito, A. M. (2012). On the recognition of emotional vocal expressions: motivations for a holistic approach. *Cogn. Process.* 13(Suppl. 2), 541–550. doi: 10.1007/s10339-012-0516-2

Esposito, A., Esposito, A. M., and Vogel, C. (2015). Needs and challenges in human–computer interaction for processing social emotional information. *Pattern Recognit. Lett.* 66, 41–51. doi: 10.1016/j.patrec.2015.02.013

Feinberg, D. R., Debruine, L. M., Jones, B. C., and Perrett, D. I. (2008). The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices. *Perception* 37, 615–623. doi: 10.1068/p5514

Feinberg, D. R., Jones, B. C., Law-Smith, M. J., Moore, F. R., Debruine, L. M., Cornwell, R. E., et al. (2006). Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice. *Horm. Behav.* 49, 215–222. doi: 10.1016/j.yhbeh.2005.07.004

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., and Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Anim. Behav.* 69, 561–568. doi: 10.1016/j.anbehav.2004.06.012

Fischer, K. (2010). Why it is interesting to investigate how people talk to computers and robots: introduction to the special issue. *J. Pragmat.* 42, 2349–2354. doi: 10.1016/j.pragma.2009.12.014

Fitch, W. T. (1999). Acoustic exaggeration of size in birds by tracheal elongation: comparative and theoretical analyses. *J. Zool.* 248, 31–49. doi: 10.1111/j.1469-7998.1999.tb01020.x

Fitch, W. T., and Kelley, J. P. (2000). Perception of vocal tract resonances by whooping cranes, Grus Americana. *Ethology* 106, 448–463. doi: 10.1046/j.1439-0310.2000.00572.x

Fitch, W. T., and Reby, D. (2001). The descended larynx is not uniquely human. *Proc. R. Soc. Biol. Sci.* 268, 1669–1675. doi: 10.1098/rspb.2001.1704

Fraccaro, P., Jones, B., Vukovic, J., Smith, F., Watkins, C., Feinberg, D., et al. (2011). Experimental evidence that women speak in a higher voice pitch to men they find attractive. *J. Evol. Psychol.* 9, 57–67. doi: 10.1556/JEP.9.2011.33.1

Fragopanagos, N., and Taylor, J. G. (2005). Emotion recognition in human–computer interaction. *Neural Netw.* 18, 389–405. doi: 10.1016/j.neunet.2005.03.006

Furui, S. (1995). Toward the ultimate synthesis/recognition system. *Proc. Natl. Acad. Sci. U.S.A.* 92, 10040–10045. doi: 10.1073/pnas.92.22.10040

Gaucher, Q., Huetz, C., Gourevitch, B., Laudanski, J., Occelli, F., and Edeline, J. M. (2013). How do auditory cortex neurons represent communication sounds? *Hear Res.* 305, 102–112. doi: 10.1016/j.heares.2013.03.011

Gerdes, A. B., Wieser, M. J., and Alpers, G. W. (2014). Emotional pictures and sounds: a review of multimodal interactions of emotion cues in multiple domains. *Front. Psychol.* 5:1351. doi: 10.3389/fpsyg.2014.01351

Gorostiza, J. F., and Salichs, M. A. (2011). End-user programming of a social robot by dialog. *Rob. Auton. Syst.* 59, 1102–1114. doi: 10.1016/j.robot.2011.07.009

Harris, T. R., Fitch, W. T., Goldstein, L., and Fashing, P. J. (2006). Black and white colobus monkey (*Colobus guereza*) roars as a source of both honest and exaggerated information about body mass. *Ethology* 112, 911–920. doi: 10.1111/j.1439-0310.2006.01247.x

Hartmann, K., Siegert, I., Gluge, S., Wendemuth, A., Kotzyba, M., and Deml, B. (2012). "Describing human emotions through mathematical modelling," in *Proceedings of the MATHMOD 2012—7th Vienna International Conference on Mathematical Modelling*, Vienna University of Technology, Vienna, Austria.

Honold, F., Bercher, P., Richter, F., Nothdurft, F., Geier, T., Barth, R., et al. (2014). "Companion-technology: towards user- and situation-adaptive functionality of technical systems," in *Intelligent Environments (IE), 2014 International Conference on*, Shanghai, 378–381. doi: 10.1109/ie.2014.60

Janak, P. H., and Tye, K. M. (2015). From circuits to behaviour in the amygdala. *Nature* 517, 284–292. doi: 10.1038/nature14188

Kim, E. S., Berkovits, L. D., Bernier, E. P., Leyzberg, D., Shic, F., Paul, R., et al. (2013). Social robots as embedded reinforcers of social behavior in children with autism. *J. Autism. Dev. Disord.* 43, 1038–1049. doi: 10.1007/s10803-012-1645-2

Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., and Elenius, K. (2011). Expression of affect in spontaneous speech: acoustic correlates and automatic detection of irritation and resignation. *Comput. Speech Lang.* 25, 84–104. doi: 10.1016/j.csl.2010.03.004

Lee, E.-J. (2010). The more humanlike, the better? How speech type and users' cognitive style affect social responses to computers. *Comput. Hum. Behav.* 26, 665–672. doi: 10.1016/j.chb.2010.01.003

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., and Paiva, A. (2013). The influence of empathy in human–robot relations. *Int. J. Hum. Comput. Stud.* 71, 250–260. doi: 10.1016/j.ijhcs.2012.09.005

Litman, D. J., and Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Commun.* 48, 559–590. doi: 10.1016/j.specom.2005.09.008

Mao, X., and Li, Z. (2010). Agent-based affective tutoring systems: a pilot study. *Comput. Educ.* 55, 202–208. doi: 10.1016/j.compedu.2010.01.005

Mauss, I. B., and Robinson, M. D. (2009). Measures of emotion: a review. *Cogn. Emot.* 23, 209–237. doi: 10.1080/02699930802204677

Mavridis, N. (2015). A review of verbal and non-verbal human–robot interactive communication. *Rob. Auton. Syst.* 63, 22–35. doi: 10.1016/j.robot.2014.09.031

Mehu, M., and Scherer, K. R. (2012). A psycho-ethological approach to social signal processing. *Cogn. Process.* 13(Suppl. 2), 397–414. doi: 10.1007/s10339-012-0435-2

Mills, T., Bunnell, H. T., and Patel, R. (2014). Towards personalized speech synthesis for augmentative and alternative communication. *Augment. Altern. Commun.* 30, 226–236. doi: 10.3109/07434618.2014.924026

Mitchell, R. L. C., Elliott, R., Barry, M., Cruttenden, A., and Woodruff, P. W. (2003). The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia* 41, 1410–1421. doi: 10.1016/S0028-3932(03)00017-4

Mitchell, W. J., Ho, C.-C., Patel, H., and Macdorman, K. F. (2011). Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management. *Comput. Hum. Behav.* 27, 402–412. doi: 10.1016/j.chb.2010.09.002

Moore, D., Mcgrath, P., and Thorpe, J. (2000). Computer-aided learning for people with autism: a framework for research and development. *Innov. Educ. Train. Int.* 37, 218–228. doi: 10.1080/13558000050138452

Mori, M. (1970). The uncanny valley. *Energy* 7, 33–35.

Morton, E. S. (1977). Occurrence and significance of motivation structural rules in some bird and mammal sounds. *Am. Nat.* 111, 855–869. doi: 10.1086/283219

Moser, E., Derntl, B., Robinson, S., Fink, B., Gur, R. C., and Grammer, K. (2007). Amygdala activation at 3T in response to human and avatar facial expressions of emotions. *J. Neurosci. Methods* 161, 126–133. doi: 10.1016/j.jneumeth.2006.10.016

Mullennix, J. W., Stern, S. E., Wilson, S. J., and Dyson, C.-L. (2003). Social perception of male and female computer synthesized speech. *Comput. Hum. Behav.* 19, 407–424. doi: 10.1016/S0747-5632(02)00081-X

Nass, C., Foehr, U. G., and Somoza, M. (2001). "The effects of emotion of voice in synthesized and recorded peech," in *Proceedings of the AAAI Emotional and Intelligent II: The Tangled Knot of Social Cognition*, (Menlo Park, CA: The AAAI Press), 91–96.

Noble, L., and Xu, Y. (2011). "Friendly speech and happy speech—are they the same?" in *Proceedings of the 17th International Congress of Phonetic Science*, Hong Kong, 1502–1505.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* 41, 1–16. doi: 10.1159/000261706

Parise, S., Kiesler, S., Sproull, L., and Waters, K. (1999). Cooperating with life-like interface agents. *Comput. Hum. Behav.* 15, 123–142. doi: 10.1016/S0747-5632(98)00035-1

Picard, R. W. (1997). *Affective Computing*. Cambridge, MA: MIT Press.

Pullin, G., and Cook, A. (2013). The value of visualizing tone of voice. *Logoped. Phoniatr. Vocol.* 38, 105–114. doi: 10.3109/14015439.2013.809144

Putze, F., and Schultz, T. (2014). Adaptive cognitive technical systems. *J. Neurosci. Methods* 234, 108–115. doi: 10.1016/j.jneumeth.2014.06.029

Reby, D., and McComb, K. (2003). Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. *Anim. Behav.* 65, 519–530. doi: 10.1006/anbe.2003.2078

Reby, D., Mccomb, K., Cargnelutti, B., Darwin, C., Fitch, W. T., and Clutton-Brock, T. (2005). Red deer stags use formants as assessment cues during intrasexual agnostic interactions. *Proc. R. Soc. B* 272, 941–947. doi: 10.1098/rspb.2004.2954

Riding, D., Lonsdale, D., and Brown, B. (2006). The effects of average fundamental frequency and variance of fundamental frequency on male vocal attractiveness to women. *J. Nonverbal Behav.* 30, 55–61. doi: 10.1007/s10919-006-0005-3

Robinson, P., and el Kaliouby, R. (2009). Computation of emotions in man and machines. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 3441–3447. doi: 10.1098/rstb.2009.0198

Ross, E. D. (2010). Cerebral localization of functions and the neurology of language: fact versus fiction or is it something else? *Neuroscientist* 16, 222–243. doi: 10.1177/1073858409349899

Rosset, D. B., Rondan, C., Da Fonseca, D., Santos, A., Assouline, B., and Deruelle, C. (2008). Typical emotion processing for cartoon but not for real faces in children with autistic spectrum disorders. *J. Autism. Dev. Disord.* 38, 919–925. doi: 10.1007/s10803-007-0465-2

Schröder, M. (2001). "Emotional speech synthesis: a review," in *Proceedings of the 7th European Conference on Speech Communication and Technology*, (Aalborg, Denmark), 561–564.

Schröder, M. (2009). "Expressive speech synthesis: past, present, and possible futures," in *Affective Information Processing*, eds J. Tao and T. Tan. (London: Springer), 111–126.

Schwenker, F., Scherer, S., and Morency, L.-P. (2015). Pattern recognition in human–computer interaction. *Pattern Recognit. Lett.* 66, 1–152. doi: 10.1016/j.patrec.2015.07.029

Siegert, I., Böck, R., and Wendemuth, A. (2013). "The influence of context knowledge for multi-modal affective annotation," in *Human–Computer Interaction. Towards Intelligent and Implicit Interaction*, ed M. Kurosu (Heidelberg: Springer Berlin), 381–390. doi: 10.1007/978-3-642-39342-6_42

Siegert, I., Böck, R., and Wendemuth, A. (2014). Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. *J. Multimodal Interfaces* 8, 17–28. doi: 10.1007/s12193-013-0129-9

Tay, B., Jung, Y., and Park, T. (2014). When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. *Comput. Hum. Behav.* 38, 75–84. doi: 10.1016/j.chb.2014.05.014

Tinwell, A., Grimshaw, M., Abdel Nabi, D., and Williams, A. (2011). Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Comput. Hum. Behav.* 27, 741–749. doi: 10.1016/j.chb.2010.10.018

Tsanas, A., Little, M. A., Mcsharry, P. E., Spielman, J., and Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *Biomed. Eng. IEEE Trans.* 59, 1264–1271. doi: 10.1109/TBME.2012.2183367

Tschacher, W., Rees, G. M., and Ramseyer, F. (2014). Nonverbal synchrony and affect in dyadic interactions. *Front. Psychol.* 5:1323. doi: 10.3389/fpsyg.2014.01323

Verma, G. K., and Tiwary, U. S. (2014). Multimodal fusion framework: a multiresolution approach for emotion classification and recognition from physiological signals. *Neuroimage* 102, 162–172. doi: 10.1016/j.neuroimage.2013.11.007

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: survey of an emerging domain. *Image Vis. Comput.* 27, 1743–1759. doi: 10.1016/j.imavis.2008.11.007

Vogeley, K., and Bente, G. (2010). "Artificial humans": psychology and neuroscience perspectives on embodiment and nonverbal communication. *Neural Netw.* 23, 1077–1090. doi: 10.1016/j.neunet.2010.06.003

Wang, Y., Hu, X., Dai, W., Zhou, J., and Kuo, T. (2014). Vocal emotion of humanoid robots: a study from brain mechanism. *Sci. World J.* 2014, 216341. doi: 10.1155/2014/216341

Wendemuth, A., and Biundo, S. (2011). "A companion technology for cognitive technical systems," in *Cognitive Behvioural Systems*, eds A. Esposito, A. M. Esposito, and A. Vinciarelli (Berlin: Springer-Verlag), 89–103.

Wood, L. J., Dautenhahn, K., Rainer, A., Robins, B., Lehmann, H., and Syrdal, D. S. (2013). Robot-mediated interviews—how effective is a humanoid robot as a tool for interviewing young children? *PLoS ONE* 8:2013. doi: 10.1371/journal.pone.0059448

Xu, Y. (2011). Speech prosody: a methodological review. *J. Speech Sci.* 1, 85–115.

Xu, Y. (2015). "Speech prosody: theories, models and analysis," in *Courses on Speech Prosody*, ed. A. R. Meireles (Newcastle: Cambridge Scholars Publishing), 146–177.

Xu, Y., Kelly, A., and Smillie, C. (2013a). "Emotional expressions as communicative signals," in *Prosody and Iconicity*, eds S. Hancil and D. Hirst (Amsterdam: John Benjamins Publishing Company), 33–60. doi: 10.1075/ill.13.02xu

Xu, Y., Lee, A., Wu, W. L., Liu, X., and Birkholz, P. (2013b). Human vocal attractiveness as signaled by body size projection. *PLoS ONE* 8:e62397. doi: 10.1371/journal.pone.0062397

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.